Learning Options using Constrained Return Variance

Arushi Jain
School of Computer Science
McGill University, Mila
Montréal, Canada
arushi.jain@mail.mcgill.ca

Doina Precup
School of Computer Science
McGill University, Mila, Google Deepmind
Montréal, Canada
dpreup@cs.mcgill.ca

Abstract

The standard setting in reinforcement learning (RL) to maximize the mean return does not assure a reliable and repeatable behavior of an agent in safety-critical applications like autonomous driving, robotics, and so forth. Often, penalization of the objective function with the variance in return is used to limit the unexpected behavior of the agent. While learning the end-to-end options have been accomplished, in this work, we introduce a novel Bellman-style direct approach to estimate the variance in return in hierarchical policies using the option-critic architecture [2]. The penalization of the mean return with the variance enables learning safer trajectories, which avoids inconsistently behaving regions. Here, we present the derivation in the policy gradient style method with the new objective function which provides the update for the option parameters in an online fashion.

1 Introduction

The objective function of maximizing the mean return does not offer any constraint on the distribution of the return, making it a vulnerable strategy for the risk-sensitive domains. The notion of avoiding risks arising from the stochastic nature of the environment (*inherent uncertainty*) using the constraint on the variance in return has been studied for a long time by the research community. [6, 9, 11, 10, 5] constrained the indirect estimate of the variance using the second-order moment methods or directly estimated the cost-to-go returns with the updates provided after completing the entire trajectory. [7] came up with a direct estimation of the variance in the λ-return using a Bellman operator in the policy evaluation methods. This work demonstrated the superiority of using the direct estimator of the variance over the indirect approaches (second-order moment methods).

Temporal abstraction provides a way to learn the policies in a hierarchical fashion which has been shown to improve exploration, robustness against model misspecification and increases the learning speed in transfer learning. [3] used the variance in the temporal difference (TD) error over the initial state-option pair distribution to estimate the controllable states in the option-critic architecture [2].

In this work, we came up with a novel hierarchical safe policy learning approach in the option-critic architecture where the hierarchical policies are learned by penalizing the direct estimate of the variance in return extending from [7] in a control setting. We seek to maximize the mean return and minimize the direct estimate of the variance in return given an initial state-option pair distribution in the policy gradient style.

2 Background

In a Markov Decision Process (MDP), an agent takes an action $a \in A$, transitions from state S_t to state S_{t+1}, and receives an immediate reward R_{t+1} from the environment. The expected reward is $r(S_t, A_t) = \sum_{r \in \mathbb{R}} r \sum_{s'} P(s', r | S_t, A_t)$ where $r : S \times A \to \mathbb{R}$. The environment dynamics is

Preprint. Under review.

modeled by $P(S_{t+1}|S_t, A_t)$, where $P : S \times A \times S \to [0, 1]$. A stochastic policy $\pi(A_t|S_t)$ determines the probability of taking an action in a given state. A MDP is represented by a tuple $\langle S, A, P, r, \gamma \rangle$, where $\gamma \in [0, 1]$ is a factor discounting the future rewards.

2.1 Option-Critic

In the option-critic architecture [2], an option $w \in W$ is defined as a tuple of $\langle I_w, \pi_w, \beta_w \rangle$; where I_w contains the initial set of states where an option can start, π_w is the option policy defining a distribution over action space and β_w determines the termination probability of an option in a state. The policy over the options is denoted by $\mu(w|s)$ describing the distribution over options given a state. Let $\Theta = [\theta, \nu, \kappa]$ be the parameters of intra-option policy π_w, termination condition β_w and policy over options μ respectively. $J_{\pi,\mu}$ denotes the objective function of maximizing the mean return. The intra-option policy gradient [2] update is:

$$\nabla_\theta J_{\pi,\mu}(\Theta) = \mathbb{E}_{\pi,\mu}[\nabla_\theta \log \pi_\theta(A_t|S_t, W_t) Q_{\pi,\mu}(S_t, W_t, A_t)],$$

and the termination gradient [2] is given by:

$$\nabla_\nu J_{\pi,\mu}(\Theta) = \mathbb{E}_{\pi,\mu}[-\nabla_\nu \beta_\nu(S_{t+1}, W_t) A_{\Theta,Q}(S_{t+1}, W_t)]$$

where, $A_{\Theta,Q}(S_{t+1}, W_t) = Q(S_{t+1}, W_t) - V(S_{t+1})$ is the advantage function describing the importance of an option value over the mean value. In the following work we assume that all the options can be started from any state ($I_w \in S, \forall\, w \in W$).

3 Safety in Option-Critic

Taking inspiration from the Bellman style equation for the variance in return introduced in the policy evaluation case in the primitive action scenario [8], we similarly derive the safe framework in the option-critic. Our notion of *safety* emphasizes minimizing the *erratic* or the *harmful* behavior of an agent in the environment [1]. The higher is the variance in return from a state; the higher would be the uncertainty in the return obtained from a state. Uncertainty in the return from a state reflects an inconsistent behavior of the agent in that particular state. Considering that the irregular or sudden behavior is classified as unsafe, potentially, the unsafe states would exhibit a higher variance in return.

Let the return be denoted by

$$G_{t,\pi,\mu} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = R_{t+1} + \gamma G_{t+1,\pi,\mu}.$$

We consider $Z_t = (S_t, W_t)$ as an augmented state space - a space of state-option pair. Here, the transition matrix over the augmented state space is given by:

$$P(z'|z, a) = P(s'|s, a)[(1 - \beta_\nu(s', w))\mathbb{1}_{w=w'} + \beta_\nu(s', w)\mu_\kappa(w'|s')] \tag{1}$$

The rewards are coming from a base MDP, where we write $r(z, a, z') = r(s, a)$. Since, $\sum_{z'} P(z'|z, a) = 1$, therefore, the reward model is defined as:

$$r(s, a) = \mathbb{E}_{\pi,\mu}[R_{t+1}|S_t = s, A_t = a] = \sum_{z'} P(z'|z, a) r(z, a, z')$$

Theorem 1. ***(Variance in Return)*** *The Bellman equation for the variance in the return from a given augmented state-action pair is:*

$$\sigma_{\pi,\mu}(z, a) = \mathbb{E}_{\pi,\mu}\big[\delta^2_{t,\pi} + \bar{\gamma}\sigma_{\pi,\mu}(Z_{t+1}, A_{t+1})\big|Z_t = z, A_t = a\big] \tag{2}$$

where $\bar{\gamma} = \gamma^2$ and δ_t is the 1-step TD error.

Proof. On expanding $G_{t,\pi,\mu} - Q_{\pi,\mu}(z, a)$,

$$\begin{aligned}
G_{t,\pi,\mu} - Q_{\pi,\mu}(z, a) =& R_{t+1} + \gamma G_{t+1,\pi,\mu} - Q_{\pi,\mu}(z, a) \\
=& R_{t+1} + \gamma \sum_{z',a'} P(z'|Z_t, A_t)\pi_\theta(a'|z') Q_{\pi,\mu}(z', a') - Q_{\pi,\mu}(z, a) \\
&+ \gamma\{G_{t+1,\pi,\mu} - \sum_{z',a'} P(z'|Z_t, A_t)\pi_\theta(a'|z') Q_{\pi,\mu}(z', a')\} \\
=& \delta_t + \gamma\Big(G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a]\Big)
\end{aligned} \tag{3}$$

Similar to [8], let the variance for the augmented state-action pair be given as:

$$\begin{aligned}
\sigma_{\pi,\mu}(z,a) &= \mathbb{E}_{\pi,\mu}\left[(G_{t,\pi,\mu} - \mathbb{E}_{\pi,\mu}[G_{t,\pi,\mu}|Z_t = z, A_t = a])^2 | Z_t = z, A_t = a\right] \\
&= \mathbb{E}_{\pi,\mu}\left[(G_{t,\pi,\mu} - Q_{\pi,\mu}(z,a))^2 | Z_t = z, A_t = a\right] \\
&= \mathbb{E}_{\pi,\mu}\Big[\Big(\delta_t + \gamma(G_{t+1,\pi,\mu} \\
&\qquad - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a])\Big)^2 | Z_t = z, A_t = a\Big] \\
&= \mathbb{E}_{\pi,\mu}\left[\delta_t^2 | Z_t = z, A_t = a\right] \\
&\quad + \gamma^2\, \mathbb{E}_{\pi,\mu}\left[(G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a])^2 | Z_t = z, A_t = a\right] \\
&\quad + 2\gamma\, \mathbb{E}_{\pi,\mu}\left[\delta_t (G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a])\right] \qquad (4)
\end{aligned}$$

Using the Lemma 1, the third term in the above (4) goes to 0. This leads the variance to

$$\sigma_{\pi,\mu}(z,a) = \mathbb{E}_{\pi,\mu}\big[\delta_{t,\pi}^2 + \bar{\gamma}\sigma_{\pi,\mu}(Z_{t+1}, A_{t+1})\big| Z_t = z, A_t = a\big].$$

□

The new safe objective function is defined as:

$$\boldsymbol{J(\Theta)} = \mathbb{E}_{z\sim d}\big[Q_\Theta(z) - \psi_z\, \sigma_\Theta(z)\big],$$

where d describes the initial state-option distribution and ψ_z is the regularizer for the variance penalty term which is a function of the augmented state space.

Theorem 2. ***(Safe Intra-Option Policy Gradient Theorem)*** *Given Markov options, $\pi_{w,\theta}$ option-policy differentiable in parameter θ, the gradient of the objective function J w.r.t. θ starting from state s and option w is:*

$$\nabla_\theta J(\Theta) = \mathbb{E}_{d,\Theta}\left[\nabla_\theta \log \pi_\theta(A_t|Z_t)\big(Q_{\pi,\mu}(Z_t, A_t) - \psi_{Z_t}\, \sigma_{\pi,\mu}(Z_t, A_t)\big)\right]$$

Proof. The gradient of the $\sigma_\Theta(z)$ w.r.t. θ is calculated in a similar fashion as the gradient of $Q_\Theta(z)$ w.r.t. θ in the *Intra-option Policy Gradient Theorem* [2]. □

The gradient update for the intra-option policy moves in the direction to maximize the mean return value and minimize the variance in return.

Theorem 3. ***(Safe Termination Gradient Theorem)*** *Given Markov options, $\beta_{w,\nu}$ termination function differentiable in parameter ν, the gradient of the objective function J w.r.t. ν starting from state s and option w is:*

$$\nabla_\nu J(\Theta) = \mathbb{E}_{d,\Theta}[-\nabla_\nu \beta_\nu(S_{t+1}, W_t)\big(A_{\pi,\mu,Q}(S_{t+1}, W_t) - \psi_z\, A_{\pi,\mu,\sigma}(S_{t+1}, W_t)\big)]$$

where $A_{\pi,\mu,\sigma}(S_{t+1}, W_t) = \sigma_\Theta(S_{t+1}, W_t) - \sigma_\Theta(S_{t+1})$ describes the difference in the variance of a given state-option pair as compared to the average scenario.

Proof. The gradient of the $\sigma_\Theta(z)$ w.r.t. ν is calculated in a similar fashion as the gradient of $Q_\Theta(z)$ w.r.t. ν in the *Termination Gradient Theorem* [2]. □

Similar to the Option-Critic, when the advantage of the value function is positive for an option, the gradient for the termination descends. On the other hand, the positive advantage function for the variance makes the gradient update for the termination ascent. It matches with the intuition, when the variance of an option is higher than the average variance, it would be desirable to terminate the option and choose a better option with a lower variance.

Theorem 4. ***(Safe Policy over Options Gradient Theorem)*** *Given Markov options, μ_κ policy over options differentiable in parameter κ, the gradient of the objective function J w.r.t. κ starting from state s and option w is:*

$$\nabla_\kappa J(\Theta) = \mathbb{E}_{d,\Theta}[\nabla_\kappa \log \mu_\kappa(W_t|S_t)\big(Q_\Theta(S_t, W_t) - \psi_{S_t,W_t}\, \sigma_\Theta(S_t, W_t)\big)]$$

when the previous option has terminated at state S_t.

Proof. Let the 1-step augmented state transition using (1) be:

$$P^{(1)}_{\bar{\gamma}}(Z_{t+1}|Z_t) = \bar{\gamma}\sum_a \pi_\theta(a|Z_t)P(Z_{t+1}|Z_t, a).$$

Similarly, the k-step transition would be defined as:

$$P^{(k)}_{\bar{\gamma}}(Z_{t+k}|Z_t) = P^{(1)}_{\bar{\gamma}}(Z_{t+k}|Z_{t+k-1}) \times P^{(k-1)}_{\bar{\gamma}}(Z_{t+k-1}|Z_t).$$

The gradient of the variance w.r.t. κ parameter following (2),

$$\begin{aligned}
\nabla_\kappa \sigma_\Theta(z) =& \nabla_\kappa \Big[\sum_a \pi_\theta(a|z)\bar{\gamma}\sum_{s'} P(s'|s,a)\big[(1-\beta_\nu(s',w))\sigma_\Theta(s',w) \\
&+ \beta_\nu(s',w)\sum_{w'}\mu_\kappa(w'|s')\sigma_\Theta(s',w')\big]\Big] \\
=& \sum_a \pi_\theta(a|z)\bar{\gamma}\sum_{s',w'} P(s'|s,a)\Big\{ \\
&\big[(1-\beta_\nu(s',w)\mathbb{1}_{w=w'} + \beta_\nu(s',w)\mu_\kappa(w'|s')]\nabla_\kappa\sigma_\Theta(s',w') \\
&+ \beta_\nu(s',w)\sum_{w'}\nabla_\kappa\mu_\kappa(w'|s')\sigma_\Theta(s',w')\Big\} \\
=& \sum_{k=0}^{\infty}\sum_{z'} P^{(k)}_{\bar{\gamma}}(z'|z)\sum_{a'}\pi_\theta(a'|z')\sum_{s''}\bar{\gamma}P(s''|s',a')\Big\{ \\
&\beta_\nu(s'',w')\sum_{w''}\nabla_\kappa\mu_\kappa(w''|s'')\sigma_\Theta(s'',w'')\Big\}
\end{aligned}$$

Similarly, the gradient of the $Q_\Theta(z)$ value function can be derived. Therefore, the update for the policy over the options parameter becomes:

$$\nabla_\kappa J(\Theta) = \mathbb{E}_{d,\Theta}[\nabla_\kappa \log\mu_\kappa(W_t|S_t)\big(Q_\Theta(S_t,W_t) - \psi_{S_t,W_t}\,\sigma_\Theta(S_t,W_t)\big)]$$

when the old option has terminated at state S_t. □

The above theorem states that the gradient of the policy over the options is also updated in the direction of maximizing the mean of return and minimizing the variance in return.

Algorithm 1 provides the pseudo-code for safe option-critic with the variance in return as the constraint.

4 Experiment

Grid-World: We experiment in the classic grid-world four rooms (FR) environment [2]. To test safety, we created a variable reward frozen patch (F) in one of the hallway generated from $\mathcal{N}(0, 8)$ distribution. The rest of the states are given a 0 reward. Agent receives a reward of 50 on reaching the goal (G) (See Fig. 1). γ is kept as 0.99. The step size of value function, variance function, intra-option policy, termination, policy over options are $0.15, 0.05, 1e-3, 5e-3, 5e-4$ respectively for both option-critic ($\psi_z = 0$) and safe option-critic ($\psi_z = 0.1$) $\forall z \in Z$.

Fig. 2a and Fig. 2b depict the performance in the FR environment. Adding the constraints on the variance in return leads to massive reduction in the standard deviation of the performance along multiple trails depicting that visits of the agent to the uncertain areas have reduced. Fig. 3 shows the sampled trajectories in the testing phase for both the baseline OC and the variance constrained Safe OC. The safe policy avoids the frozen regions in the environment. Fig. 4 portrays the variance in return induced by a policy w.r.t. different start space in the environment. The variance in return was calculated from the different 100 roll-outs (trajectory) that were played from each start state (entire state space except the goal state). This empirical variance in return was averaged over 25 different random seeds policies. The darker is the shade of the red; the higher is the value of the variance in return. This figure highlights the importance of adding the variance in return as a constraint in the standard objective function.

Algorithm 1: Safe Option-Critic with constrained return variance

Here $\alpha_x, \alpha_y, \alpha_\theta, \alpha_\kappa, \alpha_\nu$ stands for step size of critic (Q value), variance, intra-option policy, policy over options and termination respectively.
Input: a differentiable intra option policy $\pi_{\boldsymbol{\theta}}(a|s,w)$
Input: a differentiable policy over options $\mu_{\boldsymbol{\kappa}}(w|s)$
Input: a differentiable termination function $\beta_{\boldsymbol{\nu}}(s,w)$
Input: a differentiable state-action-option value $\hat{Q}_{\boldsymbol{x}}(s,a,w)$
Input: a differentiable state-action-option variance $\hat{\sigma}_{\boldsymbol{y}}(s,a,w)$
Parameters: $\gamma \in [0,1]$; variance regularizer $\psi_{s,w}$; step sizes $\alpha_x, \alpha_\theta, \alpha_y, \alpha_\nu, \alpha_\kappa$; $\alpha_\theta << \alpha_x, \alpha_y$; $\alpha_y < \alpha_x$; $\bar{\gamma} = \gamma^2$

Initialize the parameters: $\Theta = [\theta, \kappa, \nu], x, y$
$S \leftarrow s_0$; $W \leftarrow w_0$; terminated $\leftarrow$ **false**
repeat
  Observe $\{A \leftarrow \pi_\theta(.|S,W), S', R\}$
  $\delta \leftarrow R + \gamma\{(1-\beta(S',W))\hat{Q}_x(S',W) + \beta(S',W)\max_{w'}\hat{Q}_x(S',w')\} - \hat{Q}_x(S,A,W)$
  $\bar{\delta} \leftarrow \delta^2 + \bar{\gamma}\{(1-\beta(S',W))\hat{\sigma}_y(S',W) + \beta(S',W)\min_{w'}\hat{\sigma}_y(S',w')\} - \hat{\sigma}_y(S,A,W)$

  $\hat{Q}_x(S,A,W) \leftarrow \hat{Q}_x(S,A,W) + \alpha_x\delta$
  $\hat{\sigma}_y(S,A,W) \leftarrow \hat{\sigma}_y(S,A,W) + \alpha_y\bar{\delta}$
  $\theta \leftarrow \theta + \alpha_\theta \nabla_\theta \log(\pi_\theta(A|S,W))\Big(\hat{Q}_x(S,A,W) - \psi_{S,W}\,\hat{\sigma}_y(S,A,W)\Big)$
  $\nu \leftarrow \nu - \alpha_\nu \nabla_\nu \beta_\nu(S',W)\Big(\hat{A}_{\Theta,Q}(S',W) - \psi_{S',W}\,\hat{A}_{\Theta,\sigma}(S',W)\Big)$
  if terminated **then**
    $\kappa \leftarrow \kappa + \alpha_\kappa\,\nabla_\kappa \log(\mu_\kappa(W|S))\Big(\hat{Q}_x(S,W) - \psi_{S,W}\,\hat{\sigma}_y(S,W)\Big)$
  end if

  terminated $\leftarrow$ **false**
  if W terminates using $\beta_\nu(S',W)$ **then**
    $W \leftarrow \mu_\kappa(.|S')$; terminated $\leftarrow$ **true**
  end if
  $S \leftarrow S'$
until S is a terminal state

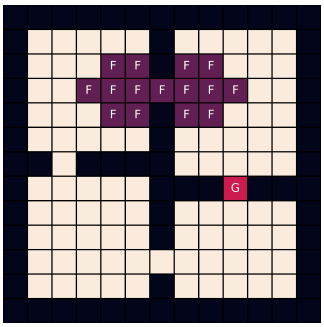
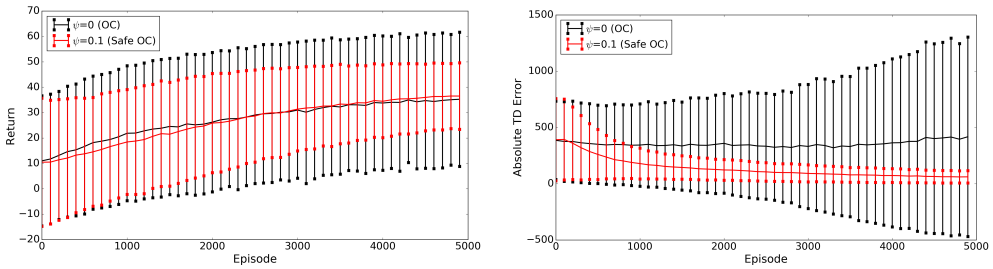


Figure 1: **Four Room (FR) Environment**: F depicts the unsafe region and G depicts the goal state.

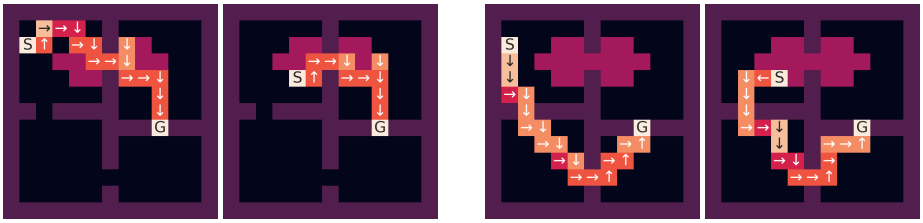
Continuous State-Action Space: Here we performed the experiments in the Mujoco environments to test the real-world use case of introducing safe trajectories while learning in an environment. We implemented our safe algorithm over existing proximal policy option-critic (PPOC) [4]. We compare the performance of the agent using both the baseline PPOC and Safe-PPOC in Fig. 5. Code for all the above experiments are added on Github[1]. The videos[2] compare the performance of the agent

[1]The Github code for the Safe-PPOC Mujoco and Discrete Grid-World experiments is here.
[2]The performance videos of the agent in PPOC and Safe PPOC in Mujoco domains is here.



(a) Learning curve (b) Absolute TD error

Figure 2: **Performance in the FR**: Shows the performance averaged over 50 trials where the vertical bands depict the std. dev. in the return across multiple trials. Shows a) the return, and b) sum of the absolute TD error. The safe policy (red) has a smaller standard deviation as compared to the baseline (black) signifying safety helps an agent to avoid variance inducing regions (unsafe).



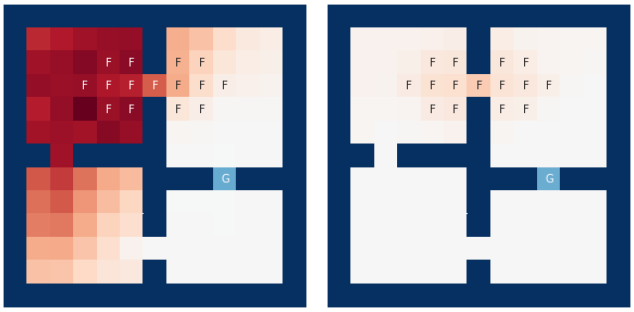
(a) OC Trajectory (b) Safe OC Trajectory

Figure 3: **Sampled Trajectories in FR environment**: a) vanilla OC and b) Safe OC with constraint variance in return. The change in the color depicts the switch among the 4 options. The safe agent takes a round about path to avoid the risky F region marked with dark purple shade in the above graphs.

using both the algorithms in the Mujoco environments. The videos highlight the effects of adding the variance in return as a constraint factor in the objective function.

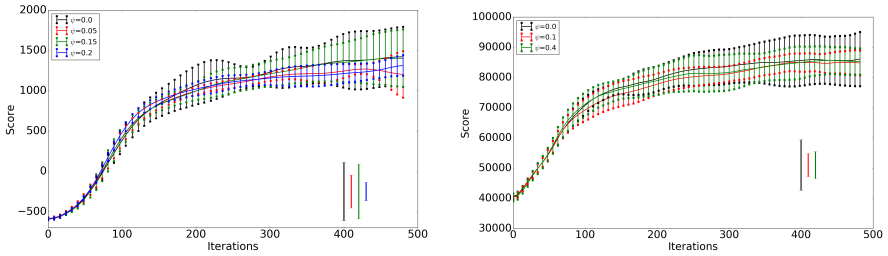
5 Conclusion & Future Work

This work aims to introduce the constraint over the variance in return to the existing option-critic architecture in order to incorporate responsible behavior in the risk-sensitive domains. Firstly, we propose a direct estimator of the variance in the hierarchical policy framework. Then, we establish a method to learn a safe and reliable policy in option-critic, which uses the above direct estimator of the variance to avoid unpredictably behaving regions. The above framework is generic, which makes no assumption about the environment, making it a simple strategy to combine with the current policy gradient techniques. The *future work* is to experiment with more different environments like Atari to understand the scalability of the safe algorithm.



(a) Variance in Baseline OC (b) Variance in Safe OC

Figure 4: **Variance in Return over trajectories w.r.t the start state**: The graph depicts the variance in return averaged over 25 different random seeds (policies) where each policy's variance in the return was calculated from 100 different roll-outs. Each cell represents the value of the variance in return when policy had the respective cell as the start state. Darker the color of the cell, higher is the variance in return. It can be observed that unsafe OC (left) has higher variance as compared to the safe OC. The diagonal lower half of (a) has lower variance because the policy does not need to pass through the unsafe F region thus lowering the variance.



(a) Half Cheetah (b) Humanoid Standup

Figure 5: **Performance in Mujoco**: Learning curve average over 10 runs where vertical bands depict the std. dev.. The vertical bars at right most corner display the std. dev. in performance over the last 50 iterations. The variance regularized PPOC ($\psi > 0$) helps in reducing the variation across multiple seed values leading to a more consistent performance.

References

[1] Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., and Mané, D. (2016). Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*.

[2] Bacon, P.-L., Harb, J., and Precup, D. (2017). The option-critic architecture. In *AAAI*, pages 1726–1734.

[3] Jain, A., Khetarpal, K., and Precup, D. (2018). Safe option-critic: Learning safety in the option-critic architecture. *arXiv preprint arXiv:1807.08060*.

[4] Klissarov, M., Bacon, P.-L., Harb, J., and Precup, D. (2017). Learnings options end-to-end for continuous action tasks. *arXiv preprint arXiv:1712.00004*.

[5] Prashanth, L. and Ghavamzadeh, M. (2013). Actor-critic algorithms for risk-sensitive MDPs. In *Advances in neural information processing systems*, pages 252–260.

[6] Sato, M., Kimura, H., and Kobayashi, S. (2001). TD algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16(3):353–362.

[7] Sherstan, C., Ashley, D. R., Bennett, B., Young, K., White, A., White, M., and Sutton, R. S. (2018a). Comparing direct and indirect temporal-difference methods for estimating the variance of the return. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 63–72.

[8] Sherstan, C., Bennett, B., Young, K., Ashley, D. R., White, A., White, M., and Sutton, R. S. (2018b). Directly estimating the variance of the λ-return using temporal-difference methods. *arXiv preprint arXiv:1801.08287*.

[9] Tamar, A., Di Castro, D., and Mannor, S. (2012). Policy gradients with variance related risk criteria. In *Proceedings of the twenty-ninth international conference on machine learning*, pages 387–396.

[10] Tamar, A., Di Castro, D., and Mannor, S. (2016). Learning the variance of the reward-to-go. *Journal of Machine Learning Research*, 17(13):1–36.

[11] Tamar, A., Xu, H., and Mannor, S. (2013). Scaling up robust MDPs by reinforcement learning. *arXiv preprint arXiv:1306.6189*.

Appendix A Proof

Lemma 1.

$$\mathbb{E}_{\pi,\mu}\left[\delta_t(G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a])\right]$$

Proof. Here, δ_t is the 1-step TD error given by:

$$\delta_t = R_{t+1} + \gamma \sum_{z',a'} P(z', a'|Z_t, A_t) Q_{\pi,\mu}(z', a') - Q_{\pi,\mu}(Z_t, A_t).$$

As δ_t is a function of (Z_t, A_t), therefore, this term can be pulled outside of the expectation.

$$\begin{aligned}
&\mathbb{E}_{\pi,\mu}\left[\delta_t(G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a])\right] \\
&\quad = \delta_t \times \mathbb{E}_{\pi,\mu}\left[G_{t+1,\pi,\mu} - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a]\right] \\
&\quad = \delta_t \times \left\{\mathbb{E}_{\pi,\mu}[G_{t+1,\pi,\mu}|Z_t = z, A_t = a] - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a]\right\} \\
&\quad = \delta_t \times \left\{\sum_{z',a'} P(z', a'|Z_t, A_t) Q_{\pi,\mu}(z', a') - \mathbb{E}_{\pi,\mu}[Q_{\pi,\mu}(Z_{t+1}, A_{t+1})|Z_t = z, A_t = a]\right\} \\
&\quad = \delta_t \times 0 = 0.
\end{aligned}$$

□